# Extracting teaching hints from a student mark prediction system in an undergraduate Computing subject

**T. D. Gedeon**

School of Computer Science & Engineering
The University of New South Wales
E-mail: tom@cse.unsw.edu.au

## Abstract

This paper describes the educationally useful knowledge extracted from the process of predicting and explaining student marks for the students in a Computing subject.

In previous work a back-propagation trained feed-forward neural network was trained to predict student performance in a large undergraduate Computer Science subject at the University of New South Wales. The prediction used continuous assessment marks from during the teaching session to predict the final grade.

The purpose of this work was to allow students to predict the final grade they are likely to achieve based on current performance, and obviously to improve their performance if the predicted grade is below their expectations. By itself, however, the network was not adequate as it provided no feedback as to why a particular student's performance merits a particular grade. We therefore generated explanations of the conclusion reached by the neural network for predicting particular student grades.

The expectation for the above process was that the predictions would be partially invalidated by student reactions, in that students with low predicted grades might work harder and so on. The knowledge extracted was sufficiently useful from the teaching perspective that the prediction system was further invalidated by the implementation of these teaching hints.

## 1 Introduction

The ability to provide general feedback on student performance and likely outcomes based on only a part of the assessment in a subject would be very useful from a number of perspectives.

From the teaching side it would be very useful to be able to identify categories of students who are not performing as well as they are able. It would be particularly desirable to be able to do this for individual students to be able to target their specific learning problems.

From a student viewpoint it would be useful to be able to determine the final mark they would receive. If the prediction was of a too low a mark, then there would be time to attempt to remedy the situation before the end of the subject. It is colloquially accepted that some students are surprised by the low marks they achieve, believing they understood the work. This can be in the face of contrary evidence in the form of continuing low continuous assessment marks. It would be harder to maintain this innocence with an actual predicted grade being available.

In the days of smaller class sizes, it was possible to do all of the above in an ad-hoc basis, relying on the experience of the teacher. With the increasing pressures on University education inexorably increasing the sizes of our lecture classes, some automated technique to aid teachers in this regard is desirable. A fully automatic process would have the added advantage that it would reduce the (perceived or true) subjectivity in a teacher produced evaluation of outcomes. Also, if the teacher provides the low evaluation, a somewhat adversarial situation is created, making remedial work a burden imposed by authority. With an automated system, the teacher is approached by students for whom the system provided low evaluations for help, and becomes an aid in their remedial process rather.

To produce an automated system, we need to reproduce the decision making of the teacher when asked to predict final performance. Unfortunately this is not well understood. Many expert tasks have been solved with good results using artificial intelligence techniques for mimicking the behaviour of experts using example data and the conclusions reached. In the next section the use of neural networks and explanation generation for this task is briefly introduced.

## 2 Background

We used a simple error back-propagation trained neural network for this task, with topology of three layers of 14 inputs, 5 hidden and 4 output neurons. The network is trained by repeatedly presenting training patterns with the 14 input values propagated through the network to the output layer, where the outputs are compared to the desired output values. Any error is used to make small modifications to the weights to decrease the error. This process is repeated recursively back to the input layer. Over a number of cycles of presentation (epochs) the network will 'learn' a generalised connection between inputs and outputs. This is validated by the use of a test set, where the input patterns are presented and propagated through the network producing network predictions of the output values without modifying the weights.

In this fashion we produced a neural network system which could reliably predict final marks given only the continuous assessment marks. This solved only half the problem, as neural networks are essentially black boxes and provide no information as to why a particular conclusion was reached. This is a general problem with the acceptability of neural networks, and is thus an active area of research.

Using our technique of causal indexes with characteristic patterns we can produce explanations for neural network conclusions on the student mark prediction problem with 94% accuracy (Turner and Gedeon, 1993, Gedeon and Turner, 1994). The causal index approach involves the calculation of the magnitude of a causal connection between all inputs and particular outputs. To avoid a combinatorial explosion, some method of reducing the magnitude of the problem is required. Our solution was to introduce the notion of characteristic patterns to serve as the context of the explanations generated. A characteristic pattern for a category is the centre of the set of patterns which cause the trained network to turn on the output neuron for that category. This has the effect that our explanations of predictions are in the context of an averaged, archetypal example as described later. In the next section we will first describe the data set used.

## 3 Data

The experiments were performed on a sample of 153 patterns taken from the class results of a second semester undergraduate Computer Science subject COMP1821 at the University of New South Wales.

```
------------------------------------------------------------------------------------------
Regno   Crse/Prog S ES   Tutgroup lab2      lab4      h2        p1        mid       final
                                       tutass    h1        lab7      f1        lab10
                                    3    5    3  20    20    3   20   20   45    3    100
-----------------------------------------------------------------------------------------
   .
0275000 3400      1 F    T10-yh     2.5  3    3   18    4.5  3   14   18.5 24    2.5  68
0275105 3420      1 F    T9-ko      3    4    2.5 17   17    3    5   14   10    2.4  56
0275139 3420      1 F    T4-ko      0    5    2.5 18   17    3    6   10   28    2.4  57
0275164 3400      1 F    T2-no      .    3    1.5  8.5  .    1.5  .    .   10.2  2.4  44
0275279 3420      1 F    T2-no      3    3    .   19   18    2    5.5  4   20    2.4  60
0275282 3400      1 F    T4-ko      2.5  3    3   19    .    3    .   10   16    2.4  51
0275298 3400      1 F    T9-ko      3    5    2.5 17   18    3    8.5 18   21    2.4  61
0275315 3420      1 F    T10-yh     2    3    0.5 14    .    1    .    .    7    2    26
0275567 3400      1 F    T10-yh     .    3.5  2.5 19.5  .    2.5  .    .   11.5  .    36
   .
```

Table 1. Raw data

The raw data consisted of the results from a number of laboratory exercises, assignments and a mid-term quiz all of which compose 40% of a student's mark for the subject. The exam mark which comprises the remaining 60% has been omitted. This omission is necessary, as a system for predicting the mark after the final examination would not be useful. The final aggregate mark is used to derive the desired output categories for the marks. The categories are:

- Distinction or higher, being: $75 \leq$ marks
- Credit, being: $65 \leq$ marks $\leq 74$
- Pass, being: $50 \leq$ marks $\leq 64$
- Fail, being: marks $\leq 49$

Note that the grade High Distinction which is 85 or above has not been separated from the Distinction category as there are relatively few of these. Also, the significance and educational relevance lies particularly in distinguishing Pass and Fail students, than between sub-categories of Distinction students.

The original set of 153 patterns is divided at random into 53 patterns to form a validation test set which will never be seen by the network during training. All data is normalised to the range 0 to 1 as is normal for neural network training and use.

## 4 Explanations

The explanations generated for the prediction of student performance is in the context of 'archetypal' cases representing the characteristics of sets of students, using a causal index technique. The causal index technique requires the calculation of the effects of input values on output values, which allows more important inputs to be distinguished as part of the explanation process.

The characteristic patterns represent the set of patterns which turn the output on for particular categories of inputs. For the four ouput categories the characteristic patterns are:

| | Crs | Stg | Enr | Tutgp | lab2 | TutAss | lab4 | H1 | H2 | lab7 | P1 | F1 | Mid | lab10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\text{CON}_{\text{Dist}}$ | 0.7 | 1 | 1 | 0.5 | 0.7 | 0.4 | 1 | 0.8 | 0.8 | 0.7 | 1 | 0.7 | 0.72 | 0.5 |
| $\text{CON}_{\text{Cred}}$ | 0 | 0.5 | 1 | 0.75 | 0.85 | 0.7 | 0.85 | 0.8 | 0.8 | 0.55 | 0.7 | 0.3 | 0.59 | 0.5 |
| $\text{CON}_{\text{Pass}}$ | 0.65 | 0.85 | 0.98 | 0.67 | 0.52 | 0.44 | 0.52 | 0.67 | 0.65 | 0.5 | 0.52 | 0.42 | 0.39 | 0.49 |
| $\text{CON}_{\text{Fail}}$ | 0.35 | 1 | 1 | 0.61 | 0.4 | 0.4 | 0 | 0.4 | 0 | 0 | 0 | 0 | 0.21 | 0 |

The mapping was made to the 0 to 1 range of the discrete values of *Crs* (Course of study, ie degree/major) *Stg* (Stage of study, ie year), *Enr* (Enrolment status, F/T or P/T), and *Tutgp* (Tutorial group identifier, there were four tutors taking two tutorials each).

Some educationally relevant knowledge can be extracted by eye from this table. The *TutAss* (Tutorial Assessment mark) is not useful for distinguishing between Pass, Fail or Distinction students. Yet this is a participation mark and is likely to be described by teachers as containing useful information. With regards the teaching, we can conclude that there is something strange about *F1* (assignment F1, functional programming) in that the values do not decrease consistently from Distinction to Fail categories.

The explanation methodology is as follows:

1 Indicate the characteristic input pattern the most similar to the input pattern.

2. Indicate ‘important’ inputs and their values in the characteristic pattern.

3. Show the set of rules obeyed.

4. Indicate the network’s next most likely output.

The first step of the explanation methodology can be compared to some forms of explanations used by human experts to explain the results they obtain. As an example consider a doctor explaining why he has come to a particular diagnosis. A typical explanation may include statements such as “You are have the classic symptoms of diabetes.” Presenting the characteristic input pattern is similar to this kind of behaviour, while listing of the specific symptoms and their values are similar to our notion of indicating the important inputs and their characteristic values. The doctor's explanation of how the symptoms and values such as "blood sugar of over 8" together indicate the diagnosis is again similar to our set of rules obeyed.

## 5 Information Extracted

The explanations for the conclusion for a particular student is in the context of the most similar characteristic pattern, for each output. For the Distinction result we will describe the process in detail for each of the possible characteristic pattern. That is, the output may be a Distinction, but the student behavious may be characteristic of a Distinction student, a Credit student, and so on.

### 5.1 Rules for Dist

#### 5.1.1 Distinction result with input pattern similar to $CON_{Dist}$

The simplest case occurs when the input pattern is most similar to the Characteristic ON pattern for the same output. Thus, the input pattern looks like a classic Distinction case, and is categorised as such by the trained network.

The Causal Index graph for the $CON_{Dist}$ pattern is shown in Figure 1.

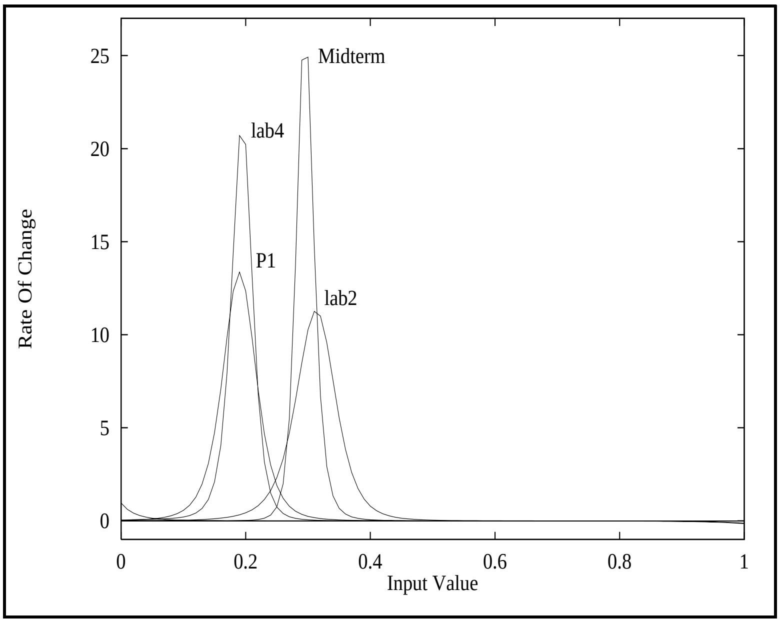

Figure 1. Causal Index of inputs for Mark Predictor network – $CON_{Dist}$

The four positive peaks in the graph produce simple numerical rules, combined using *AND* operators, as discussed previously. These rules derived from this pattern are shown below, in Table 2. The rules derived from the *Characteristic ON* pattern will be called ‘standard’ rules.

| Characteristic Pattern | Rule Set |
|---|---|
| $CON_{Dist}$ | (lab2 ≥ 0.44) AND (lab4 ≥ 0.23) AND<br>(P1 ≥ 0.27) AND (Midterm ≥ 0.37) |

Table 2. Rules from the Causal Index graph for Mark Predictor network – $CON_{Dist}$

The interpretation of these rules in the context of the example needs some discussion. The rules would be presented to the user after presenting the most similar characteristic pattern.

The rules in Table 2 indicate the performance required so as to actually match the Distinction archetype respresented by $CON_{Dist}$. For example, so long as the second laboratory assessment (*lab2*) mark is not below *0.44*, the result will be a Distinction mark for a particular input pattern representing a particular student's performance. This explains the quite low values of the inputs in the rule set compared with the values in the characteristic pattern.

The characteristic pattern indicates that fairly high marks are required overall to achieve a Distinction result. The four inputs identified in the rule set in Table 2 are very plausible to be so significant. There are two possible explanations for this, firstly the weighting of these four in the final grade is high, or secondly, the material covered in those four assessments is representative of the material in the final exam. While the weighting of *Midterm* is high, it is also clearly representative of the material in the final examination.

The other three inputs identified as important are not weighted particularly high in the calculation of the final grade. Lack of a good level of performance during laboratory exercises and the procedural programming assignment plausibly shows a deficiency leading to a less than Distinction performance.

A major strand of the subject is on functional programming, and thus it is initially surprising that the assignment *F1* does not occur in the rule set. This strand is taught at a more introductory level, and may be either sufficiently well understood in general, or not understood at all, that a particularly low performance in the assignment does not modify the final grade. Note that this accords with our previous observation that the values for *F1* in the various characteristic patterns is surprising.

### 5.1.2 Distinction result with input pattern similar to $CON_{Cred}$

In this case the most similar characteristic pattern is not the Distinction pattern. However, the trained network on the particular input pattern decides that the categorisation is Distinction by turning that output unit on. Thus, the input pattern looks like a classic Credit case, but is categorised as a Distinction by the trained network.

There are no cases of this nature. All patterns in our set which are most similar to the Credit mark's *Characteristic ON* pattern are actually either categorised as a Credit mark by the trained network, or the pattern satisfies the unmodified rules for a Distinction result derived from the $CON_{Dist}$ pattern.

### 5.1.3 Distinction result with input pattern similar to $CON_{Pass}$

In this case the input pattern is most similar to the *Characteristic* pattern corresponding to the Pass output. Note that the actual categorisation produced by the trained neural network was a Distinction, notwithstanding the closer similarity of the pattern to that of a standard Pass student. We will extract rules here to explain the categorisation in terms of the difference from the relevant standard patterns. Note also that the *Characteristic* pattern $CON_{Pass}$ is referred to as $COFF_{Dist}^{Pass}$ when it is being used as an *OFF* pattern for the *Distinction* output. The Causal Index graph (not shown) shows that a change in a particular single variable can produce a Distinction result from an otherwise Pass result.

The rules produced are shown in Table 3.

| Characteristic Pattern | Rule Set |
|---|---|
| $\mathrm{COFF}_{\mathrm{Dist}}^{\mathrm{Pass}}$ | Midterm ≥ 0.85 |

Table 3. Rules produced from the Causal Index graph for Mark Predictor network – $COFF_{Dist}^{Pass}$

This rule set is actually formed by modifying the rule set for the *Characteristic ON* pattern of the Pass output with the rule extracted for the case of a Distinction decision. Since there is only one term, it is completely replaced by the modified form.

What we have found from the extracted rule, is that there is a population of students who perform at a passing level in continuous assessment, but do particularly well in the Midterm examination and get a

Distinction grade. We can readily accept that a good result in the Midterm will correlate with a good result in the final examination. Further work is required to discover how we can differentiate 'real' Pass students from these students during the semester.

#### 5.1.4 Distinction result with input pattern similar to $CON_{Fail}$

Unsurprisingly, there are no cases of this nature.

### 5.2 Rules for Cred

The rules for Credit pattern of marks with a Credit result are shown below in Table 4.

| Characteristic Pattern | Rule Set |
|---|---|
| $\mathrm{CON}_{\mathrm{Cred}}$ | (Course < 0.35) AND (Stage < 0.93) AND (lab2 ≥ 0.39) AND (Midterm ≥ 0.12) AND (lab10 ≥ 0.14) |

Table 4. Rules from the Causal Index graph for Mark Predictor network – $CON_{Cred}$

The interpretation of these rules in the context of the most similar characteristic pattern which is the $CON_{Cred}$, shown in Table 1.

The rules in Table 4 indicate the performance required so as to actually match the Credit archetype respsrented by $CON_{Cred}$. For example, so long as the second laboratory (*lab2*) mark is above *0.39*, the result will be a *Credit* mark for the particular input pattern representing a particular student's performance, even though the expected value would be *0.7*, which is the value of *lab2* in the *Characteristic ON* pattern.

The characteristic pattern indicates that quite high marks are required overall to achieve a Credit result. The Course, and Stage inputs are more unusual. Students who are not in the first year stage of their degree in general only do this subject if they have failed it in the past and are repeating, and hence unlikely to get a Credit for the subject. The encoding of the Course code gives low values to Science/Engineering majors, and higher values towards more Humanities based majors. Thus, a student whose performance is similar to the characteristic Credit pattern is more likely to achieve a Credit mark if he or she is enrolled in a scientific major.

Note also that students with performance similar to the Credit archetype will receive a Credit (assuming the other conditions hold) with quite low Midterm marks.

For Credit result with a pattern of performance characteristic of a Distinction there are the following rules shown in Table 11, below:

| Characteristic Pattern | Rule Set |
|---|---|
| $\mathrm{COFF}^{\mathrm{Dist}}_{\mathrm{Cred}}$ | (lab2 ≥ 0.38) AND (lab4 ≥ 0.23) AND (P1 ≥ 0.27) AND (Midterm ≥ 0.37) AND (Course ≥ 0.92) |

Table 5. Rules produced from the Causal Index graph for Mark Predictor network – $COFF^{Dist}_{Cred}$

Note that the rule set is formed from the Distinction rule set augmented by the single term for the student's course code value.

The course code input (*Course*) has an unfortunate encoding, with low values corresponding to Business and Information Technology Students, medium to low values corresponding to various Arts and Social Science courses, to medium to high values for students enrolled in Science/Law and miscellaneous courses, with the highest values for the Science & Mathematics course. Our method was able to extract useful information even with this encoding, as we will discuss in a later section. The modification of the Distinction rule set selects out the Science & Mathematics majors. This implies that of students showing a characteristically Distinction performance, those enrolled in this major will most likely still reduce their overall performance in the final examination to a Credit level. This is surprising, and may be due to the relative (low) significance of a Computer Science subject to their major or their performance. That is, students in Science who do this (terminating) subject may be lower performers and doing it as a filler, while the non-Science students may tend to be among the more enterprising, higher performing students in their disciplines.

There are no examples of Credit results from students with characteristic performance of either Pass or Fail.

### 5.3 Rules for Pass

A single rule can be extracted. This is shown in Table ?.

| Characteristic Pattern | Rule Set |
|---|---|
| $CON_{Pass}$ | Midterm < 0.75 |

Table 6. Rule from the Causal Index graph for Mark Predictor network – $CON_{Pass}$

The single rule from Table ? seems too all encompassing. This is only if we take it out of context. In the context of an explanation for patterns which are most similar to the characteristic pattern for the Pass output shown below, it makes a lot more sense.

| | Crs | Stg | Enr | Tutgp | lab2 | TutAss | lab4 | H1 | H2 | lab7 | P1 | F1 | Mid | lab10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $CON_{Pass}$ | 0.65 | 0.85 | 0.98 | 0.67 | 0.52 | 0.44 | 0.52 | 0.67 | 0.65 | 0.5 | 0.52 | 0.42 | 0.39 | 0.49 |

That is, from all of the patterns which are similar to the classic Pass pattern, those which obey the rule are Pass results. Thus

Clearly those which do not obey the rule above, and are similar to the classic Pass pattern must belong to some other class and obey some of the rules found elsewhere in this paper. Note that our method is 94% accurate in explanations, thus some 6% of cases are not covered by our rules.

For students whose performance is characteristic of a Distinction student but get a pass, there are a number of ways to achieve this. The rules as for Distinction students with substitution of one or more of the clauses below. Thus, really low marks in two minor assessments, or being in the bottom 30% of the class in the midterm reduces a Distinction to a Pass.

| Characteristic Pattern | Rule Set |
|---|---|
| $COFF_{Pass}^{Dist}$ | (lab4 < 0.03) OR (P1 < 0.05) OR (Midterm < 0.3) |

Table 7. Rules from the Causal Index graph for Mark Predictor network – $COFF_{Pass}^{Dist}$

Similarly there are an even larger number of ways to convert a performance characteristic of a Credit student to a pass.

| Characteristic Pattern | Rule Set |
|---|---|
| $COFF_{Pass}^{Cred}$ | ((Course ≥ 0.51) AND (Course < 0.63)) OR<br>(Stage ≥ 0.99) OR (lab2 < 0.22) OR<br>(lab10 < 0.01) OR (Midterm < 0.02) |

Table 8. Rules from the Causal Index graph for Mark Predictor network – $COFF_{Pass}^{Cred}$

The values of the *Course* input selected indicate that the students from some of the miscellanous non-science based courses choosing this Computing subject, result in Pass grades even when they otherwise perform to a Credit level as judged by pattern similarity. The very high value for *Stage* denotes later year students exclusively. This is probably due to a preponderance of students in later years being repeat students who previously failed. This is probably because continuous assessment is more likely to be similar to previous iterations of the course than the final exam. The very low marks on *lab10* and the *Midterm* exam are straightforward, as is the significance of the low mark for *lab2* indicating a Pass instead of a Credit result.

One may also conjecture somewhat from the gap between the rule as discussed earlier for Pass outputs (*Midterm* $< 0.75$ produces a Pass grade), and the result for Distinction results (*Midterm* $\geq 0.85$ produces a Distinction grade). It seems plausible that a value of *Midterm* between these values would predispose to a Credit result. Unfortunately, such intermediate values do not as a single change produce a Credit result.

Surprisingly, there are no single changes which will produce a Pass result if the most similar pattern was the Characteristic ON pattern for the Fail output. From an educational viewpoint this is heartening, indicating the robustness of the Fail results, in that one single improvement in performance would not have boosted a Fail mark to a Pass.

## 5.4 Rules for Fail

The rules indicating when the pattern similarity to the standard *Fail* pattern does mean a *Fail* result are shown in Table ?.

| Characteristic Pattern | Rule Set |
|---|---|
| $\mathrm{CON}_{\mathrm{Fail}}$ | (H2 < 0.52) AND (Midterm < 0.98) |

Table 9. Rules from the Causal Index graph for Mark Predictor network – $CON_{Fail}$

The term referring to the *Midterm* input is easiest to explain. Any pattern which is most similar to the standard *Fail* pattern will be a *Fail* result so long as the mark in the *Midterm* examination is less than *0.98*. Further, the mark in the *H2* assignment must be (approximately) less than average. Clearly, an extremely high mark in the *Midterm* would not correlate with an overall *Fail*. The *H2* assignment is the last assignment, and a better than average result could only occur at that late stage in the semester by a significant improvement of the student’s efforts and so on. This would then lead on to a better result in the final examination.

For students whose result is a Fail, but otherwise performed at a Pass level, the rule extracted is shown in Table ?.

| Characteristic Pattern | Rule Set |
|---|---|
| $\mathrm{CON}_{\mathrm{Fail}}^{\mathrm{Pass}}$ | H2 ≥ 0.58 |

Table 10. Rules from the Causal Index graph for Mark Predictor network – $COFF_{Fail}^{Pass}$

This rule indicates that for any student whose profile is generally closest to a *Pass*, it is not a good idea to spend too much time on assignment *H2*. This is comprehensible, since the last assignment is still not worth a lot of marks, so some of the time spent getting a good mark would have been better invested into studying for the final examination.

# 6 Teaching hints and conclusions

The most significant information extracted was that not very high pass marks on the *H2* assignment was a likely concommitant of some Pass students failing overall. In retrospect this can be explained on the basis that Computing students do tend to spend too much time on programming tasks to the detriment of their examination studying.

The other major hint was the absence of *F1* from all the rule sets, and the non-monotonic behaviour of a major strand of the subject between the categories of students from Distinction to Fail. The conclusion on closer examination was that there was a problem with the nature of the assignment. That is, there were two main ways to solve the problem. The brute force method was very time consuming, but produced good results eventually. The better structured, more elegant solution was actually a lot harder than it looked. Thus, Pass students would opt for the easy option and spend more time detailing a solution and get a good mark, while Credit students would attempt to use the more elegant solution and only get it partially working and score lower marks. The Distinction students would make the elegant solution work.

The subject was of course modified using these hints derived from the performance data. The predition system is still not available to students as the subject keeps changing due to this kind of information, and external influences.

## References

Gedeon, TD and Turner, H “Extracting Contextual if-then Rules from a Feedforward Neural Network,” *Proceedings Brazil-Japan Joint Symposium on Fuzzy Systems*, 10 pages, Manaus, 1994.

Turner, H and Gedeon, TD “Extracting Meaning from Neural Networks,” *Proceedings 13th International Conference on AI*, vol. 1, pp. 243-252, Avignon, 1993.